

HIV Synonymous and Nonsynonymous Substitution Frequencies

1. HIV-1 and HIV-2 *gag* and *env* substitution frequencies.

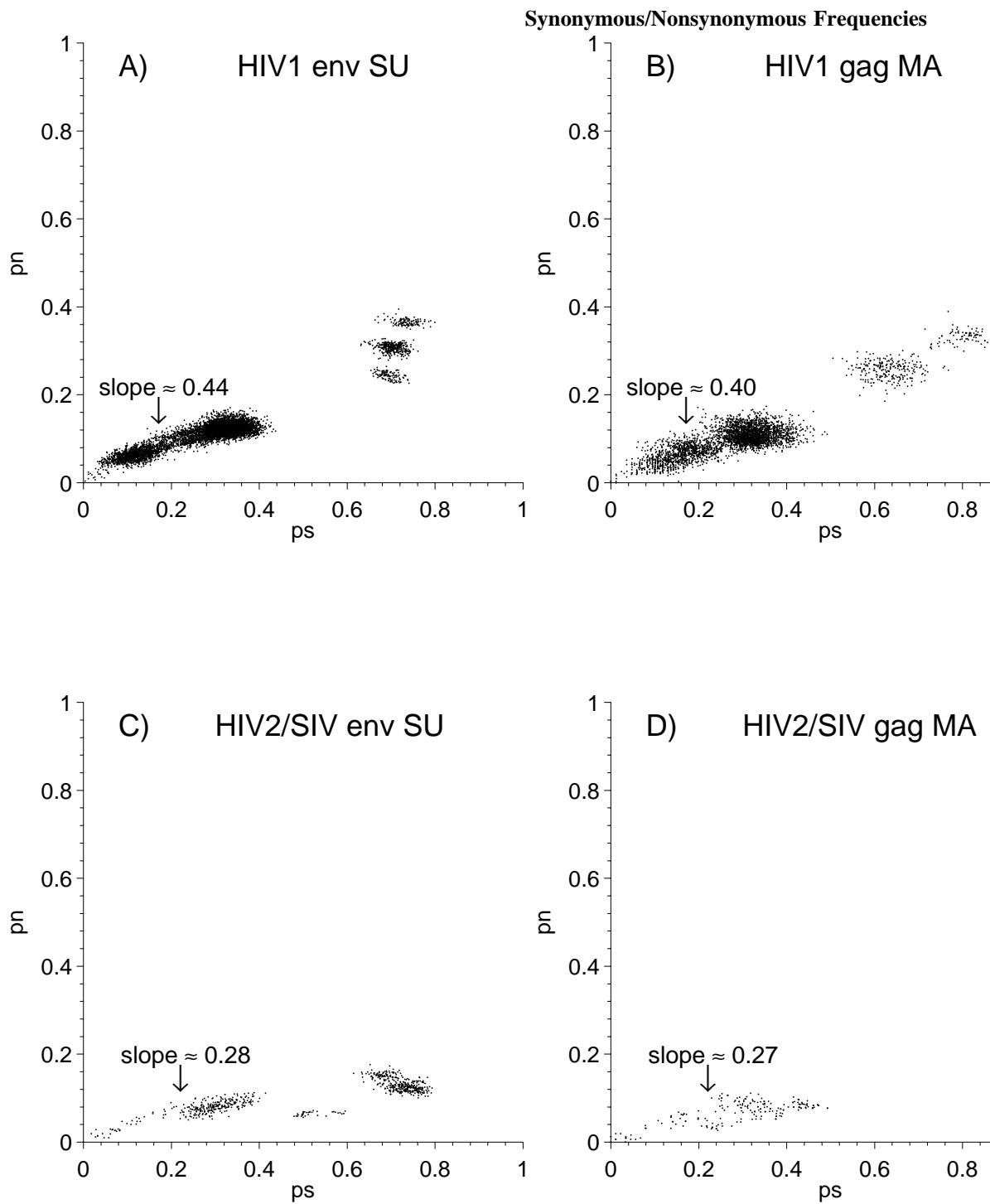
Of the possible kinds of nucleotide substitutions in a typical viral coding sequence, synonymous (or so-called “silent”) substitutions are the most commonly observed. If “mutational saturation” has not been reached for any of the pairwise relationships in a homologous set of sequences, synonymous substitution frequencies can provide a linear measure of genetic variation under conditions of minimal selection pressure. Nonsynonymous substitutions (generating amino acid replacements) tell us something about negative and positive Darwinian selection; under certain circumstances, they also serve as a metric. We should keep in mind the possibility that subtle selective pressures may be exerted upon synonymous substitutions (we have seen in the previous section that base composition is a factor in this regard), as well as the possibility that amino acid replacements may be unselected (Kimura’s neutral theory). The majority of sites in a coding sequence will be nonsynonymous targets, however the majority of changes observed will be synonymous, because these produce the least negative effects.

Nonsynonymous (pn) and synonymous (ps) substitution frequencies are shown in the accompanying plots for a large number of pairwise relationships over a) HIV-1 *env* SU, b) HIV-1 *gag* MA, c) HIV-2/SIV *env* SU, and d) HIV-2/SIV *gag* MA coding sequences. The Nei-Gojobori algorithm was used to determine these frequencies [1]. “Initial” slopes have been calculated, using a least-squares program, over the range of ps and pn values of 0.0 to 0.3; multiple hits are likely to be encountered for values greater than 0.3. Synonymous substitution frequencies approach saturation in the range of 0.6 to 0.8; the saturation levels for nonsynonymous substitutions are less certain due to the very nature of positive selection effects.

Of note in these plots is first, the uncanny similarity of *env* SU slopes to *gag* MA slopes—0.44 and 0.40 for HIV-1s, 0.28 and 0.27 for HIV-2s. Equivalently measured slopes for *gag* CA coding sequences (not shown)—0.15 for HIV-1s and 0.09 for HIV-2/SIVs—clearly indicate that the near-identities needn’t have been the case. We don’t have an explanation for this phenomenon at this time: it seems to suggest that either *env* SU and *gag* MA are “co-selected” in some sense, or that they are both demonstrating some fixed level of neutral mutation not seen in *gag* CA coding sequences. Synonymous substitution frequencies at the scale of complete coding sequences are generally uniform across the viral genomic molecule; that is to say no gradient of intrinsic base substitution has been observed for HIVs.

A second finding is that HIV-1s possess higher slopes of pn to ps than do HIV2/SIVs for *env* and *gag* coding sequences. This result is consistent with a study conducted by Shpaer and Mullins [2], which was based upon a smaller data set and a different method for determining substitution frequencies.

Finally, the extent of synonymous changes are approximately equivalent for the majority of HIV-1, HIV-2, and SIV sequence relationships—ps of 0.4—although this cannot be taken as evidence of identical ages of the sequence sets. The clusters of more mutated sequence relationships, ps greater than 0.4, come about through comparison of HIV-1 M group sequences to O group sequences, and through comparison of HIV-2 sequences of one subtype to those of another subtype and to SIVs. The former effect—the nearly identical extents of synonymous substitution in HIV-1s and HIV-2s—is probably merely a consequence of sampling: sequences tend to be reported for individuals who have been recently infected and therefore haven’t diverged that much from other recently infected (and sampled) sources.



Synonymous/Nonsynonymous Frequencies

2. HIV-1 Intersubtype Comparisons

As new HIV-1 *env* coding sequences have become available, phylogenetic tree analyses (III-C) have pointed toward a “star” relationship among the M group of HIV-1s and separately among the O group of HIV-1s [3]. An analysis of synonymous and nonsynonymous substitution frequencies provides some corroboration of the hypothesis that HIV-1s are part of two star phylogenies, which are suggestive of two separate introductions of virus at about the same time into the human population.

In the following analysis, 340 nt from the C2-V3 coding regions of HIV-1 subtypes A through H (of the M group) were analyzed using the Nei-Gojobori algorithm [1]. The average frequencies of ps (synonymous substitutions) and pn (nonsynonymous substitutions) were determined for members of a given sequence subtype (e.g., A) in relation to all other subtype sequences (e.g., B through H). Ideally, such a study would involve equal numbers of each subtype; however, that is not possible with the current data sets. Notwithstanding this limitation, the averages are remarkably identical as shown in the schematic summary below. Preliminary results based upon unpublished O group sequences (not shown) point to highly similar averages among these viral sequences—moreover, values that are similar to the average value shared by group M subtype sequences [3].

Average intersubtype synonymous and nonsynonymous substitution frequencies are given for each subtype set relative to all other subtype sequences; thus ‘.38/.18’ gives the average ps and pn values respectively for A group sequences. Although these averages were calculated over a shorter coding stretch and involved more sequences, they are comparable to the ones reported in refs 1 and 3 (slightly higher in every case as is to be expected for C2-V3 relative to *env* gp120, hence average pn to ps ratios closer to 0.5 than to 0.4, as shown in the previous section).

References

- [1] Korber et al., *J. Virol.* **68**:6730–6744, 1994.
- [2] Shpaer, G and Mullins, J., *J. Mol. Evol.* **37**:57–65, 1993).
- [3] Myers, *AIDS Res. Hum. Retroviruses* **10**:1317–1323, 1994.

